Assessing Statistical Significance of Overrepresented Oligonucleotides

Alain Denise, Mireille Regnier, Mathias Vandenbogaert

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Assessing statistical significance of overrepresented oligonucleotides

Alain DENISE, Mireille REGNIER, Mathias VANDENBOGAERT

N ° 4132
February 2001

THÈME 2

ISSN 0249-6399 ISRN INRIA/RR--4132--FR+ENG

Assessing statistical significance of overrepresented oligonucleotides

Alain DENISE, Mireille REGNIER, Mathias VANDENBOGAERT

Thème 2 — Génie logiciel
et calcul symbolique
Projet Algo

Rapport de recherche n ° 4132 — February 2001 — 16 pages

Abstract: Assessing statistical significance of overrepresentation of exceptional words is becoming an important task in computational biology. We show on two problems how large deviation methodology applies. First, when some oligomer H occurs more often than expected, e.g. may be overrepresented, large deviations allow for a very efficient computation of the so-called p-value. The second problem we address is the possible changes in the oligomers distribution induced by the overrepresentation of some pattern. Discarding this noise allows for the detection of weaker signals. Related algorithmic and complexity issues are discussed and compared to previous results. The approach is illustrated with two typical examples of applications on biological data.

(Résumé : tsvp)

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : (33) 01 39 63 55 11 – Télécopie : (33) 01 39 63 53 30

Signification statistique de la surrepresentation d'oligonucléotides

Résumé : La signification statistique de la surrepresentation de mots exceptionnels est un sujet important en bioinformatique. Nous Utilisons la théorie des grandes déviations sur deux problèmes. En premier, nous considerons un motif H surrepresenté. Les grandes déviations permettent un calcul très efficace de la queue de la distribution, la p-valeur des biologistes. En second, nous considérons les changements dans la distribution des mots induite par la surrepresentation d'un mot donné. La suppression de ce bruit permet alors la detection de signaux plus faibles. Les conséquences algorithmiques associés et la complexité sont comparés aux résultats existants. Cette approche est illustrée sur des données réelles dans deux exemples biologiques typiques.

Assessing statistical significance of overrepresented oligonucleotides*

Alain Denise[†‡] Mireille Régnier[§¶]

Mathias Vandenbogaert[§‖]

Abstract

Assessing statistical significance of overrepresentation of exceptional words is becoming an important task in computational biology. We show on two problems how large deviation methodology applies. First, when some oligomer H occurs more often than expected, e.g. may be overrepresented, large deviations allow for a very efficient computation of the so-called p-value. The second problem we address is the possible changes in the oligomers distribution induced by the overrepresentation of some pattern. Discarding this noise allows for the detection of weaker signals. Related algorithmic and complexity issues are discussed and compared to previous results. The approach is illustrated with two typical examples of applications on biological data.

Keywords: Promoter sequences, coregulation, motif statistics, large deviations, p-values.

1 Introduction

Putative DNA recognition sites can be defined in terms of an idealized sequence that represents the bases most often present at each position. Conservation of only very short consensus sequences is a typical feature of regulatory sites (such as promoters) in both prokaryotic and eukaryotic genomes.

*This research was partially supported by IST Program of the EU under contract number 99-14186 (ALCOM-FT), REMAG Action from INRIA and IMPG French program.

‖LaBRI, Université Bordeaux I, 351, Cours de la Liberation, 33405 Talence, France

Structural genes are often organized into clusters that include genes coding for proteins whose functions are related. It is common for the genes coding for the enzymes of a metabolic pathway to be organized into such a cluster. Other related activities may be included in the unit of coordinated control. A set of coregulated genes may contain similar promoter sequences, so that, when needed, the expression of the corresponding proteins is activated by the interaction of a common transcription factor, the so-called sigma factor. There exists more than one type of sigma factor, each specific for a different class of promoter. Changes in sigma factors appear in some cases when there is a wholesale reorganization of transcription, in response to general environmental changes (e.g. heat shock genes). Each sigma factor causes RNA polymerase to initiate at a particular set of promoters, so that, in vivo, recognition occurs efficiently only in the presence of the appropriate sigma-factor, and so that transcription of different groups of enzymes is mutually exclusive. Research is very active in this area [GK97, vHACV98, RHEC98, RVD98, VMS99, VGMMdA00, PMG00, KBCC00, BFW$^+$00]. In these works, one searches for exceptional patterns in nucleotidic sequences, using various tools to assess the significance of such rare events.

Large deviation is a mathematical area that deals with rare events; to our knowledge, it has not really been used in computational biology. Nevertheless, our recent results in [DR01], that extend preliminary results in [RS97] show it may be a very powerful method to assess statistical significance of very rare events. We discuss this approach on two biological examples found in published papers.

The first problem we address is the following. One considers a *candidate*, e.g. a word that occurs more often than expected. One needs to quantify this difference between the observation and the expectation. Among the classical statistical tools, the so-called p-values are much more precise than the Z- scores (or the χ-scores). The drawback is that their computation is considered as much harder. Large deviations provide a very efficient way to compute them in some cases.

As a second problem, we consider some consequences of the overrepresentation of a word on a sequence distribution . In particular, it has been observed that, whenever a word is overrepresented, its subwords or the words that contain it, look overrepresented. Such words are called below *artefacts* [BFW$^+$00]. It is a desirable goal to choose the best element in the set composed of a word and its artefacts. It is also important to discard automatically the "noise" created by the artefacts, in order to detect other words that are

potentially overrepresented. An important example is the noise introduced by the *Alu* sequences. An other one is the χ-sequence `GNTGGTGG` in *H. influenzae* [Nic00]. We provide some mathematical results and the algorithmic consequences.

Efficiency of this approach comes from the existence of explicit formulae for the (conditioned) distribution. Large deviations allow for a very fast computation. Moreover, due to the "simplicity" of the result -if not of the proof-, their implementation is easy and provides numerically stable and guaranteed computations. Hence, they occasionnally correct commonly used approximations. Still, computing the correct result is much faster and precise than computing the approximation. Approach is valid for various counting models. For a sake of clarity, we present it for the most commonly used, the overlapping model [Wat95].

In Section 4 and Section 5, we validate our approach by a comparison with published results derived by other methods that are computationally more expensive. In Section 6, we discuss possible improvements and present further work.

2 Statistical tools in computational biology

Our aim in this section is not a formal and exhaustive discussion. We rather remind basic useful definitions for statistical criteria. We briefly discuss their limits, e.g the validity domains and the computational efficiency. Below, we denote by $O(\mathrm{H})$ the number of observations of a given pattern H in a given sequence. Depending of the application, it may be either the number of occurrences [PMG00, KBCC00] or the number of sequences where it appears [BFW$^+$00, vHACV98].

Z-scores Many definitions of this parameter can be found in the literature. Other names can be used: see for instance the so-called *contrast* used in [PMG00]. A common feature is that they compare the observation with the expectation and the variance. A rather general definition is

$$Z(\mathrm{H}) = \frac{E(H) - O(\mathrm{H})}{\sqrt{V(H)}} \tag{1}$$

where H is a given pattern or word, $O(\mathrm{H})$ is the *observed* number of occurrences, $E(H)$ the expectation and $V(H)$ the variance. Many recent works

allow for a fast computation of E and V, hence Z. Relevant approximations are discussed in [RLM00], notably the Poisson approximation $V = E$. Nevertheless, if Z-scores are a very efficient filter to detect potential candidates, they are not precise enough. Notably, this parameter is not stable enough for very exceptional words, e.g. when the expectation is much smaller than 1. This will be detailed in Section 4. Moreover, it is relevant only for large sequences, and does not adapt easily to the search in several small sequences.

p-values For each word that occurs r times in a sequence or in a set of N (related) sequences, one computes the probability that this event occurs just "by chance":

$$pval(\mathrm{H}) = P(O(\mathrm{H}) \geq r) \ . \tag{2}$$

When the expectation of a given word is much smaller than 1, a single occurrence is a rare event. In this case, the p-value is defined as: $P(O(\mathrm{H}) \geq r$ knowing that $O(\mathrm{H}) \geq 1)$, e.g.:

$$pval(\mathrm{H}) = \frac{P(O(\mathrm{H}) \geq r)}{P(O(\mathrm{H}) \geq 1)} \ . \tag{3}$$

The computation is performed in two steps. The probability that H occurs in a given sequence is known. An exact formula is provided in [RS97] and used in [KBCC00]. An approximated formula is often used, for instance in [BFW$^+$00] or in software RSA-tools (http://copan.cifn.unam.mx/ jvanheld/rsa-tools/). Then, a binomial formula provides (2). It may be approximated [BFW$^+$00] by the incomplete β-function. Nevertheless, any computation is rather delicate, and machine dependent as numerical stability necessitates a very careful use of real precision.

3 Main results

3.1 Basic notations

The model of random text that we handle with is the *Bernoulli model*: one assumes the text to be randomly generated by a memoryless source. Each letter s of the alphabet has a given probability p_s to be generated at any step. Generally, the p_s are not equal.

Definition 3.1 *Given a pattern* H *of length* m *on the alphabet* $\mathcal{S}$ *and a Bernoulli distribution on the letters of* $\mathcal{S}$*, the probability of* H *is defined as*

$$P(\mathrm{H}) = \prod_{i=1}^{m} p_{\mathrm{H_i}}$$

where $p_{\mathrm{H_i}}$ *denotes the* i*-th character of* H*. By convention, empty string* ϵ *has probability* 1*.*

Finding a pattern in a random text is, in some sense, correlated to the previous occurrences of the same or other patterns [PBM91]. Hence for example, the probability of finding $\mathrm{H}_1 = \texttt{ATT}$ knowing that one has just found $\mathrm{H}_2 = \texttt{TAT}$ is - intuitively - rather good since a `T` right after H_2 is enough to give H_1. *Correlation polynomials* and *correlation functions* give a way to formalize this intuition.

Definition 3.2 *The* correlation set *of two patterns* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *is the set of words* w *which satisfy: there exists a non-empty suffix* v *of* $\mathrm{H_i}$ *such that* $vw = \mathrm{H_j}$*. It is denoted* $\mathcal{A}_{i,j}$*. If* $\mathrm{H_i} = \mathrm{H_j}$*, then the correlation set is called the* autocorrelation set *of* $\mathrm{H_i}$*.*

Thus for example, the correlation set of $\mathrm{H}_1 = \texttt{ATT}$ and $\mathrm{H}_2 = \texttt{TAT}$ is $\mathcal{A}_{1,2} = \{AT\}$; the autocorrelation set of H_1 is $\{\epsilon\}$, while the autocorrelation set of H_2 is $\{\epsilon, \texttt{AT}\}$. Empty string always belong to the autocorrelation set of any pattern.

Definition 3.3 *The* correlation polynomial *of two patterns* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *of length* m_i *and* m_j *is defined as:*

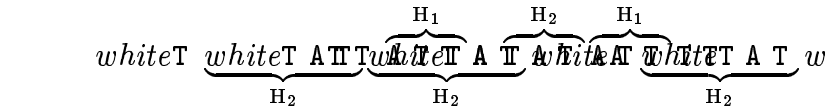
$$A_{i,j}(z) = \sum_{w \in \mathcal{A}_{i,j}} P(w) z^{|w|} \ ,$$

where $|w|$ *denotes the length of word* w*. If* $\mathrm{H_i} = \mathrm{H_j}$*, then this polynomial is called the* autocorrelation polynomial *of* $\mathrm{H_i}$*. The* correlation function *is:*

$$D_{i,j}(z) = (1-z)A_{i,j}(z) + P(\mathrm{H_j}) z^{m_j} \ .$$

. When $H_i = H_j$*, the correlation function can be written* D_i*.*

The most common counting model is the *overlapping model*: overlapping occurrences of patterns are taken into account. It is as follows. For example, consider two oligonucleotides $\mathrm{H}_1 = \mathtt{ATT}, \mathrm{H}_2 = \mathtt{TAT}$ and a sequence TTATTATATATT. Sequence contains 2 occurrences of H_1 and 4 occurrences of H_2, as shown below:



It turns out [DR01] that our main results rely on the computation of the (real) roots of a polynomial equation:

Definition 3.4 *Let a be a real number such that $a > P(\mathrm{H}_1)$. Let (E_a) be the* fundamental equation*:*

$$D_1(z)^2 - (1 + (a-1)z)D_1(z) - az(1-z)D_1'(z) = 0 \ . \tag{4}$$

Let z_a be the largest real positive solution of Equation (E_a) that satisfies $0 < z_a < 1$. z_a is called the fundamental root *of (E_a).*

3.2 p-value for a single pattern

Main result of this section is the theorem below, that provides the probability for the observed number of occurrences to be much greater than the expectation.

Theorem 3.1 *Let H_1 be a given pattern, and k be its observed number of occurrences in a random sequence of length n. Denote $a = \frac{k}{n}$ and assume that $a > P(\mathrm{H}_1)$. Then:*

$$pval(\mathrm{H}_1) = Prob(O(\mathrm{H}_1) \geq k) \approx \frac{1}{2\sigma_a\sqrt{n}} e^{-nI(a)} \tag{5}$$

where

$$I(a) = a \ln\left(\frac{D_1(z_a)}{D_1(z_a) + z_a - 1}\right) + \ln z_a \ , \tag{6}$$

$$\sigma_a^2 = a(a-1) - a^2 z_a \left(\frac{2D_1'(z_a)}{D_1(z_a)} - \frac{(1-z_a)D_1''(z_a)}{D_1(z_a) + (1-z_a)D_1'(z_a)}\right) \tag{7}$$

and z_a is the fundamental root of E_a. $I(a)$ is called the rate function. *Additionnally:*

$$Prob(O(\mathrm{H}_1) = k) \approx \frac{1}{\sigma_a\sqrt{2\pi n}} e^{-nI(a)} \quad . \tag{8}$$

Remark: When $a = P(\mathrm{H}_1)$, the number of H_1-occurences is equal to its expected value. Conditional variance σ_a in (7) becomes: $\sigma = P(\mathrm{H}_1)(2A_1(1) - 1 + (1-2m)P(\mathrm{H}_1))$, e.g. the unconditional variance computed by various authors [Wat95, RSW00].

Remark: The two probabilities $Prob(O(\mathrm{H}_1) \geq k)$ and $Prob(O(\mathrm{H}_1) = k)$ appear to be very similar in magnitude. Formulae above provide an attractive alternative to the incomplete β-function, as they are easier to program, faster and much more stable numerically.

3.3 Conditioning by an overrepresented word

In this subsection, we assume a pattern H_1 has been detected as an overrepresented word and we provide mathematical results to investigate the changes induced on the sequence distribution. Intuitively, the artefacts of an overrepresented word should look overrepresented. For example, if $\mathrm{H}_1 = AATAAA$, any word $\mathrm{H2} = ATAAAN$ is an artefact. A rough approximation of its expected value is $O(\mathrm{H}_1) \times \frac{P(N)}{P(A)}$. As $O(\mathrm{H}_1) >> E(\mathrm{H}_1)$, this is much greater than unconditioned expectation $E(\mathrm{H}_1) \times \frac{P(N)}{P(A)}$. The theorem below, proven in [DR01], establishes the precise formulae:

Theorem 3.2 *Given two patterns H_1 and H_2, assume the number of H_1-occurrences, $O(\mathrm{H}_1)$, is known and equal to k, with $a = \frac{k}{n} \geq P(\mathrm{H}_1)$. Then, the conditional expectation of $O(\mathrm{H}_2)$ is:*

$$E(O(\mathrm{H}_2)/O(\mathrm{H}_1) = k) \approx n\alpha \tag{9}$$

where α is a function of a, autocorrelation functions and probabilities:

$$\alpha = a\frac{D_{1,2}(z_a) \times D_{2,1}(z_a)}{D_1(z_a)(D_1(z_a) + z_a - 1)} \tag{10}$$

and z_a is the fundamental root of Equation (4).

*Remark:*In the central region, e.g. $k = nP(\mathrm{H}_1)$, substitutions $a = P(\mathrm{H}_1)$ and $z_a = 1$ in (9) yield $\alpha = P(\mathrm{H}_2)$, if H_1 and H_2 do not overlap.

Once a dominating signal has been detected, one looks for a weaker signal by a comparison of the number of observed occurrences of patterns with their conditional expectations. This procedure automatically eliminates artefacts. An example is provided in Section 5. It also allows for a choice of the best candidate between a word and its artefacts.

Computational complexity Another approach is used in *Regexpcount* [Nic00]. Although our formal proof of Theorem 3.2 relies on similar mathematical tools, our *explicit formulae* allow for skipping the expensive intermediate computations (bivariate generating functions, recurrences,...), hence provide a much faster algorithm.

4 Tandem repeats in B. subtilis and A. thaliana

In [KBCC00], authors search for localized repeats with a statistical filter. Software *Except* relies on a simple basic idea: long approximate repeats are likely to contain multiple exact occurrences of shorter words. DNA sequences are divided into overlapping fragments of size n. This size n is a parameter of the algorithm chosen for each run. Typically, n ranges from 250 to 5000. In each window, the p-value is computed for any pattern that occurs more than once. As the total number of occurences remains relatively small (typically 3 to 5), exact computation through generating functions is (theoretically) possible. Nevertheless, this approach, chosen by the authors, is computationally expensive. Typically, r repeated multiplications of polynomials of degree n. This gives a time complexity $O(n \log n \log r)$, if a Fast Fourier Transform is used, and numerical stability is rather delicate.

Large deviation computation for rare events reduces to the numerical computation of real roots of a polynomial equation. This is efficiently implemented in *Maple*. Results are given in Figure 4 for one 2008 nucleotides long fragment in *A. thaliana* where 5 approximate tandem repeats of a 40-uple were found. For each oligonucleotide, the first value is the number of occurrences in the window, the second is the probability $P(\mathrm{H})$, the third one is the p-value computed by our large deviation formulae. The fourth one is the p-value computed in [KBCC00] with a generating function method and the last one is the Z-score. We notice that, for any pattern, the p-values computed with two different methods are of the same magnitude order . However, then can differ up to a factor 1.72. This can be due to a combina-

Oligomer	Obs.	Probability	p-value (large dev).	p-value[KBCC00]	Z-score
AAGACGGTT	3	1.876×10^{-6}	2.186×10^{-6}	2.780×10^{-6}	48.95
AATTGGCGG	2	8.428×10^{-7}	8.059×10^{-4}	8.343×10^{-4}	48.71
ACGACGCTT	4	1.455×10^{-6}	1.604×10^{-9}	0.982×10^{-9}	74.01
ACGCTTGG	4	1.107×10^{-6}	5.374×10^{-10}	4.391×10^{-10}	84.93
ACGGTTCAC	3	1.455×10^{-6}	2.265×10^{-6}	1.458×10^{-6}	55.49
GAGAAGACG	5	5.487×10^{-7}	0.687×10^{-14}	1.180×10^{-14}	151.10
TTTGTACCA	3	8.430×10^{-6}	4.350×10^{-5}	4.611×10^{-5}	22.96

Figure 1: Measures on the 7 oligonucleotides considered in [KBCC00]

tion of several factors, like the approximation done in our calculations and the possible numerical instability of computations in [KBCC00].

On the other hand, the last column of the table confirms that Z-score is not adequate for very rare events. Patterns AAGACGGTT and AATTGGCGG have the same Z-score 48, while p-values have a ratio 100. For patterns ACGACGCTT and ACGCTTGG, the two parameters define a different order. The same inversion appears between AATTGGCGG and TTTGTACCA.

5 Polyadenylation signals in human genes

In [BFW+00], Beaudoing et al. study polyadenylation signals in mRNAs of human genes. One of their aims is to find several variants of the well known AAUAAA signal. For this purpose, they select 5646 putative mRNA 3' ends of length 50 nucleotides and seek for overrepresented hexamers. Pattern AAUAAA is clearly the most represented: it occurs in 3286 sequences, for a total number of 3456 occurrences. Seeking for other (weaker) signals involves searching for other overrepresented hexanucleotides. Nevertheless, it is necessary to avoid *artefacts*, e.g. patterns that appear overrepresented because they are similar to the first pattern. The algorithm designed by Beaudoing et al. consists in cancelling all sequences where the overrepresented hexamer has been found. Hence, they search for the most represented hexamer in the 2780 sequences which do not contain the strong signal AAUAAA.

Here we show how Theorem 3.2 gives a procedure for dropping the artefacts of a given pattern without cancelling the sequences where it appears. Figure 5 presents the 15 most represented hexamers in the sequences consid-

Hexamer	Obs.	Rank	Exp.	Z-sc.	Rank	Cond.Exp.	Cond.Z-sc.	Rank
AAUAAA	3456	1	363.16	167.03	1			**1**
AAAUAA	1721	2	363.16	71.25	2	1678.53	1.04	1300
AUAAAA	1530	3	363.16	61.23	3	1311.03	6.05	404
UUUUUU	1105	4	416.36	33.75	8	373.30	37.87	**2**
AUAAAU	1043	5	373.23	34.67	6	1529.15	-12.43	4078
AAAAUA	1019	6	363.16	34.41	7	848.76	5.84	420
UAAAAU	1017	7	373.23	33.32	9	780.18	8.48	211
AUUAAA	1013	8	373.23	33.12	10	385.85	31.93	**3**
AUAAAG	972	9	184.27	58.03	4	593.90	15.51	34
UAAUAA	922	10	373.23	28.41	13	1233.24	-8.86	4034
UAAAAA	922	11	363.16	29.32	12	922.67	9.79	155
UUAAAA	863	12	373.23	25.35	15	374.81	25.21	**4**
CAAUAA	847	13	185.59	48.55	5	613.24	9.44	167
AAAAAA	841	14	353.37	25.94	14	496.38	15.47	36
UAAAUA	805	15	373.23	22.35	21	1143.73	-10.02	4068

Figure 2: Table of the most frequent hexanucleotides. *Obs*: number of observed occurrences. *Exp.*: (non-conditional) expectation. *Cond.Exp.*: expectation conditioned by number of occurrences of AAUAAA.

ered in [BFW+00]. Columns 2 and 3 respectively give the observed number of occurrences and the rank according to this criteria. Columns 4, 5 and 6 present the (non-conditioned) expected number of occurrences, the corresponding Z-score and the rank of the hexamer according to this Z-score. Here, the variance has been approximated by the expectation; this is possible as stated in [RLM00]. Remark that rankings of columns 3 and 6 are quite similar: only patterns UAAAAA and UAAAUA do not belong to both rankings. A number of motifs look like the canonical one: they may be artefacts. This is confirmed by the three last columns which present, respectively, the expected number of occurrences conditioned by the observed number of occurrences of AAUAAA, the corresponding conditioned Z-score and the rank according to this criteria. It is clear that artefacts are dropped out, generally very far away in the ranking. It is worth noticing that some patterns which seemed overrepresented are actually avoided: this is the case for AUAAAU which goes down from 5th to last place (among the 4096 possible hexamers, only 4078

are present in the sequences). As AUAAAU is an artefact of the strong signal, this means that U is rather avoided right after this signal.

The case of UUUUUU in rank 2 is particular: this pattern is effectively overrepresented, but was not considered by Beaudoing et al. as a putative polyadenylation signal because its position does not match with observed requirements (around -15/-16 nucleotides upstream of the putative polyadenylation site.) It should also be stated that the approximation of the variance by the expectation that we do for all patterns is not as good for periodic patterns like UUUUUU as for others [RLM00]. By this way, variance of UUUUUU is underevaluated; so its actual Z-score is significantly lower than the one given in the table.

Now overrepresentation of AUUAAA (rank 3) is obvious; this is the known first variant of the canonical pattern.

We remark that the following hexamer, UUAAAA, is an artefact of AUUAAA. It suggests to define a conditional expectation, or, even better, a p-value that takes into account the overrepresentation of two or more signals instead of one: in this example, AAUAAA *and* AUUAAA. This extension of Theorem 3.2 is the subject of a future work.

6 Conclusion and Perspectives

In this paper, we illustrated a possible use of large deviation methods in computational biology. These results allow, in some cases, a very fast computation of p-values that is numerically stable. These preliminary results are quite appealing and should be extended in several directions. First, it may be necessary to eliminate several strong independent signals [BFW$^+$00]. A second task is the simplification of our formulae for artefacts: this would allow to achieve automatically the choice between a word and its subwords. A third task is the extension to the computation of the p-value for a set of small sequences. Finally, regulatory sites may also be associated with structured motifs [MS00] and extension to this case should be realized.

References

[BFW$^+$00] E. Beaudoing, S. Freier, J. Wyatt, J.M. Claverie, and D. Gautheret. Patterns of Variant Polyadenylation Signal Usage in

Human Genes. *Genome Research.*, 10:1001–1010, 2000.

[DR01] A. Denise and M. Régnier. Word statistics conditioned by overrepresented words, 2001. in preparation;http://algo.inria.fr/regnier/index.html.

[GK97] M.S. Gelfand and E.V. Koonin. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymess. *Nucleic Acids Research*, 25(12):2430–2439, 1997.

[KBCC00] Maude Klaerr-Blanchard, Hélène Chiapello, and Eivind Coward. Detecting localized repeats in genomic sequences: A new strategy and its application to *B. subtilis* and *A. thaliana* sequences. *Comput. Chem.*, 24(1):57–70, 2000.

[MS00] L. Marsan and M.F. Sagot. Extracting structured motifs using a suffix tree-algorithms and application to promoter consensus identification. In *RECOMB'00*, pages 210–219. ACM-, 2000. Proceedings RECOMB'00, Tokyo.

[Nic00] P. Nicodème. The symbolic package Regexpcount. In *GCB'00*, 2000. presented at GCB'00, Heidelberg, October 2000; available at http://algo.inria.fr/libraries/software.html.

[PBM91] P.A. Pevzner, M. Borodovski, and A. Mironov. Linguistic of Nucleotide sequences:The Significance of Deviations from the Mean: Statistical Characteristics and Prediction of the Frequency of Occurrences of Words. *J. Biomol. Struct. Dynam.*, 6:1013–1026, 1991.

[PMG00] E.M. Panina, A.A. Mironov, and M.S. Gelfand. Statistical analysis of Complete Bacterial Genomes:Avoidance of Palindromes and Restriction-Modification Systems. *Genomics. Proteomics. Bioinformatics*, 34(2):215–221, 2000.

[RHEC98] F.R. Roth, J.D. Hughes, P.E. Estep, and G.M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.*, 16:939–945, 1998.

[RLM00] M. Régnier, A. Lifanov, and V. Makeev. Three variations on word counting. In *GCB'00*, pages 75–82. Logos-Verlag, 2000. Proc. German Conference on Bioinformatics, Heidelberg; submitted to BioInformatics.

[RS97] M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4):631–649, 1997. preliminary draft at ISIT'97.

[RSW00] G. Reinert, S. Schbath, and M. Waterman. Probabilistic and Statistical Properties of Words: An Overview. *Journal of Computational Biology*, 7(1):1–46, 2000.

[RVD98] E. Rocha, A. Viari, and A. Danchin. Oligonucleotides bias in *bacillus subtilis*: general trands and taxonomic comparisons. *Nucl. Acids Research*, 26:2971–2980, 1998.

[VGMMdA00] A.T. Vasconcelos, M.A. Grivet-Mattoso-Maia, and D.F. de Almeida. Short interrupted palindromes on the extragenic DNA of Escherichia coli K-12, Haemophilus influenzae and Neisseria meningitidis. *BioInformatics*, 16(11):968–977, 2000.

[vHACV98] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.

[VMS99] A. Vanet, L. Marsan, and M.-F. Sagot. Promoter sequences and algorithmical methods for identifying them. *Res. Microbiol.*, 150:779–799, 1999.

[Wat95] M. Waterman. *Introduction to Computational Biology*. Chapman and Hall, London, 1995.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105,
78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS
Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
(France)
http://www.inria.fr
ISSN 0249-6399